Exploratory Data Analysis Report

Titanic Dataset

Author: [Pooja Yashwant Dhaware]
Date: December 16, 2025
Course/Project: Task 5 - Exploratory Data Analysis
Tools Used: Python, Pandas, Matplotlib, Seaborn

Executive Summary

This report presents a comprehensive exploratory data analysis of the Titanic dataset, which contains information about 891 passengers aboard the RMS Titanic. The analysis aims to uncover patterns, trends, and insights related to passenger survival rates using visual and statistical methods.

Key Finding: Survival was strongly influenced by gender, passenger class, and fare paid, with females and first-class passengers having significantly higher survival rates.

Table of Contents

1. Introduction

Background

The RMS Titanic sank on April 15, 1912, after colliding with an iceberg during its maiden voyage. This tragedy resulted in the loss of over 1,500 lives, making it one of the deadliest maritime disasters in history.

Objective

The objective of this analysis is to:

- Understand the demographic composition of passengers
- Identify factors that influenced survival rates
- Discover patterns and relationships within the data
- Provide insights through visual and statistical exploration

Methodology

This analysis employs descriptive statistics and data visualization techniques using Python libraries including Pandas, Matplotlib, and Seaborn.

---

2. Dataset Overview

Dataset Description

The Titanic dataset contains passenger information with the following characteristics:

Dataset Dimensions:

- Total Records: 891 passengers
- Total Features: 15 columns
- Target Variable: Survived (0 = No, 1 = Yes)

Feature Categories:

1. **Demographic:** Age, Sex
2. **Socioeconomic:** Pclass (Passenger Class), Fare
3. **Family:** SibSp (Siblings/Spouses), Parch (Parents/Children)
4. **Travel:** Embarked (Port of Embarkation), Cabin
5. **Outcome:** Survived

Data Dictionary

Column	Description	Data Type
survived	Survival (0 = No, 1 = Yes)	int
pclass	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)	int
sex	Gender	object
age	Age in years	float
sibsp	Number of siblings/spouses aboard	int
parch	Number of parents/children aboard	int
fare	Passenger fare	float
embarked	Port of embarkation (C=Cherbourg, Q=Queenstown, S=Southampton)	object
class	Passenger class	category
who	Man, woman, or child	object
deck	Deck location	category

3. Data Quality Assessment

3.1 Missing Values Analysis

Missing Data Summary:

Feature	Missing Count	Percentage
deck	688	77.2%
age	177	19.9%
embark_town	2	0.2%
embarked	2	0.2%

Observations:

- The 'deck' variable has the highest missing rate (77.2%), limiting its usefulness for analysis
- Age has approximately 20% missing values, which is significant but manageable
- Embarked location has minimal missing data (only 2 records)
- Other variables have complete data

Visualization: [INSERT: missing_values.png]

3.2 Data Types and Structure

Numerical Variables: age, fare, sibsp, parch, survived, pclass
Categorical Variables: sex, embarked, class, who, deck, embark_town, alive

All data types are appropriate for their respective variables.

---

4. Univariate Analysis

4.1 Target Variable: Survival Rate

Overall Survival Statistics:

- Survived: 342 passengers (38.4%)
- Died: 549 passengers (61.6%)

Observation: The majority of passengers (61.6%) did not survive the disaster, highlighting the severity of the tragedy.

Visualization: [INSERT: survival_rate.png]

4.2 Age Distribution

Statistical Summary:

- Mean Age: 29.7 years
- Median Age: 28.0 years
- Standard Deviation: 14.5 years
- Age Range: 0.42 to 80 years

Observations:

- The age distribution is relatively normal with a slight right skew

- Most passengers were between 20-40 years old
- There were passengers of all ages, including infants and elderly
- Several outliers exist in the upper age range (70+ years)

Visualizations: [INSERT: age_distribution.png]

4.3 Fare Distribution

Statistical Summary:

- Mean Fare: $32.20
- Median Fare: $14.45
- Standard Deviation: $49.69
- Fare Range: $0 to $512.33

Observations:

- The fare distribution is highly right-skewed with extreme outliers
- The median fare ($14.45) is much lower than the mean ($32.20), indicating the presence of high-value outliers
- Most passengers paid relatively low fares (under $50)
- A small number of passengers paid premium fares (over $200)

Visualizations: [INSERT: fare_distribution.png]

4.4 Passenger Class Distribution

Distribution:

- Third Class: 491 passengers (55.1%)
- First Class: 216 passengers (24.2%)
- Second Class: 184 passengers (20.7%)

Observation: More than half of the passengers traveled in third class, reflecting the ship's capacity to accommodate large numbers of immigrants and lower-income travelers.

Visualization: [INSERT: class_distribution.png]

4.5 Gender Distribution

Distribution:

- Male: 577 passengers (64.8%)
- Female: 314 passengers (35.2%)

Observation: Males significantly outnumbered females, comprising nearly two-thirds of all passengers.

Visualization: [INSERT: gender_distribution.png]

---

5. Bivariate Analysis

5.1 Survival by Gender

Survival Rates:

- Female Survival Rate: 74.2%
- Male Survival Rate: 18.9%
- Difference: 55.3 percentage points

Observations:

- Females had dramatically higher survival rates than males
- This aligns with the "women and children first" evacuation protocol
- Gender was one of the strongest predictors of survival

Visualization: [INSERT: survival_by_gender.png]

5.2 Survival by Passenger Class

Survival Rates by Class:

- First Class: 62.96%
- Second Class: 47.28%
- Third Class: 24.24%

Observations:

- Clear inverse relationship between class number and survival rate
- First-class passengers had 2.6x higher survival rate than third-class passengers
- This suggests socioeconomic status significantly impacted survival chances

- Possible factors: cabin location (proximity to lifeboats), priority access, better awareness of danger

Visualization: [INSERT: survival_by_class.png]

5.3 Age vs Survival

Observations:

- Children (age 0-10) had relatively higher survival rates
- Young adults (20-40) had mixed survival outcomes
- Elderly passengers (60+) had lower survival rates
- Age distribution between survivors and non-survivors shows significant overlap

Statistical Insight:

- Mean age of survivors: 28.3 years
- Mean age of non-survivors: 30.6 years
- Difference: 2.3 years (minimal difference)

Visualization: [INSERT: age_vs_survival.png]

5.4 Fare vs Survival

Observations:

- Passengers who survived generally paid higher fares
- Median fare for survivors: $26.00
- Median fare for non-survivors: $10.50
- This correlation likely reflects the relationship between fare and passenger class

Visualization: [INSERT: fare_vs_survival.png]

5.5 Age vs Fare Relationship

Observations:

- Weak positive correlation between age and fare
- Higher fares were paid across all age groups
- No strong age-based pricing pattern evident

- Fare was more strongly associated with class than age

Visualization: [INSERT: age_vs_fare_scatter.png]

---

6. Multivariate Analysis

6.1 Correlation Analysis

Key Correlations with Survival:

1. **Fare (0.257):** Positive correlation - higher fares associated with better survival
2. **Pclass (-0.338):** Negative correlation - lower class numbers (higher class) associated with better survival
3. **Sex:** Strong categorical relationship (not shown in numeric correlation)

Other Notable Correlations:

- Pclass and Fare: -0.549 (strong negative - higher class = higher fare)
- Age and Pclass: -0.369 (older passengers in higher classes)
- SibSp and Parch: 0.415 (families traveled together)

Visualization: [INSERT: correlation_heatmap.png]

6.2 Survival by Class and Gender

Combined Analysis:

Class	Female Survival	Male Survival	Difference
First	96.8%	36.9%	59.9%
Second	92.1%	15.7%	76.4%
Third	50.0%	13.5%	36.5%

Observations:

- Female passengers had higher survival rates across ALL classes
- The gender advantage was strongest in second class (76.4% difference)
- Even third-class females had better survival rates than first-class males

- This confirms that gender was the strongest predictor, followed by class

Visualization: [INSERT: survival_class_gender.png]

6.3 Pair Plot Analysis

Key Relationships Identified:

- Pclass and Fare: Strong inverse relationship
- Age distribution: Similar across survival outcomes
- SibSp/Parch: Small families had slight survival advantage
- No strong linear relationships between most numerical variables

Visualization: [INSERT: pairplot.png]

---

7. Key Findings & Insights

7.1 Primary Findings

1. Gender was the Strongest Survival Predictor

- Females had 74.2% survival rate vs. 18.9% for males (55.3% difference)
- "Women and children first" protocol was clearly followed
- This held true across all passenger classes

2. Socioeconomic Class Significantly Impacted Survival

- First-class passengers: 63.0% survival rate
- Third-class passengers: 24.2% survival rate
- Higher class passengers had better cabin locations and priority access to lifeboats

3. Fare Correlated with Survival

- Higher fares were associated with better survival rates
- This relationship is largely explained by the correlation between fare and passenger class
- Premium accommodations provided advantages during evacuation

4. Age Had Moderate Impact

- Children had priority in evacuation
- No dramatic difference in mean age between survivors and non-survivors

- Age alone was not as predictive as gender or class

5. Family Size Considerations

- Traveling alone slightly reduced survival chances
- Very large families (5+ members) also had lower survival rates
- Small families (2-4 members) had optimal survival rates

7.2 Patterns and Trends

Demographic Patterns:

- The ship carried predominantly male, third-class passengers
- Age distribution was relatively normal with most passengers in their 20s-30s
- Most passengers embarked from Southampton (72%)

Survival Patterns:

- Clear hierarchy: First-class females > Second-class females > Third-class females > First-class males > Second/Third-class males
- Children prioritized in evacuation
- Crew members had relatively low survival rates

Economic Patterns:

- Strong correlation between fare paid and passenger class
- Wide disparity in fare prices ($0 to $512)
- Most passengers paid modest fares under $50

7.3 Anomalies and Outliers

Identified Anomalies:

1. Some passengers paid $0 fare (possibly crew or special cases)
2. Extreme high fares (over $500) for luxury suites
3. A few third-class passengers had relatively high survival rates
4. Some first-class males did not survive despite advantages

Missing Data Considerations:

- 77% missing cabin information limits spatial analysis
- 20% missing age data may slightly bias age-related findings

- Minimal impact from other missing values

---

8. Conclusions & Recommendations

8.1 Conclusions

This exploratory data analysis of the Titanic dataset reveals that survival was not random but strongly influenced by demographic and socioeconomic factors:

1. **Gender was paramount:** The "women and children first" maritime protocol was followed, resulting in dramatically different survival rates between males and females.

2. **Class discrimination existed:** Passenger class significantly affected survival chances, with first-class passengers having 2.6x better odds than third-class passengers.

3. **Multiple factors interacted:** The combination of being female AND in first class provided the highest survival probability (96.8%), while being male in third class had the lowest (13.5%).

4. **Economic barriers mattered:** Higher fares corresponded to better survival rates, reflecting both cabin location and evacuation priority.

8.2 Historical Context

These findings align with historical accounts of the Titanic disaster:

- Limited lifeboats (enough for only ~1,178 of 2,224 people)
- Evacuation protocol prioritizing women and children
- Third-class passengers faced locked gates and were located farther from lifeboats
- First-class passengers had better access to information and resources

8.3 Data Science Insights

For Predictive Modeling:

- Gender and Pclass should be primary features
- Fare can serve as a proxy for class when class data is missing
- Age should be included but has less predictive power
- Family size (derived from SibSp + Parch) may be useful

Feature Engineering Opportunities:

- Create "Title" from name (Mr., Mrs., Miss., Master.)
- Binary "IsChild" feature (age < 18)
- "FamilySize" = SibSp + Parch + 1
- "IsAlone" binary feature
- "FareGroup" categorical binning

8.4 Recommendations for Further Analysis

1. **Text Analysis:** Extract and analyze passenger names for titles and social status
2. **Cabin Analysis:** Investigate the 23% with cabin data for location-based patterns
3. **Network Analysis:** Examine family and social connections
4. **Predictive Modeling:** Build classification models (Logistic Regression, Random Forest, XGBoost)
5. **Time-Series:** If embarkation times available, analyze temporal patterns

---

9. Appendix

9.1 Technical Details

Environment:

- Python 3.x
- Jupyter Notebook
- Libraries: pandas 1.x, matplotlib 3.x, seaborn 0.11.x, numpy 1.x

Data Source:

- Seaborn built-in dataset
- Alternative: Kaggle Titanic Competition dataset

9.2 Code Repository

All code used for this analysis is available in the accompanying Jupyter notebook: `Titanic_EDA.ipynb`

9.3 Statistical Tests Performed

- Descriptive statistics (mean, median, mode, std dev)
- Correlation analysis (Pearson correlation)
- Value counts and frequency distributions
- Group-by aggregations for survival rates

9.4 Visualization Summary

Total visualizations created: 15+

- Distribution plots: 6
- Categorical plots: 5
- Correlation analysis: 1
- Multivariate plots: 3

9.5 Limitations

1. Missing cabin data (77%) limits spatial analysis
2. Missing age data (20%) introduces potential bias
3. Dataset represents only survivors who provided information
4. Crew member data is limited
5. No time-based information for sequence of events

9.6 References

- Kaggle Titanic Dataset: https://www.kaggle.com/c/titanic
- Seaborn Documentation: https://seaborn.pydata.org/
- Pandas Documentation: https://pandas.pydata.org/
- Historical Records: Encyclopedia Titanica

Acknowledgments

This analysis was conducted as part of Task 5: Exploratory Data Analysis. Special thanks to the open-source community for providing excellent data science tools and the Kaggle community for maintaining the Titanic dataset.

End of Report

dhawarep04@gmail.com